

COMPARISON AND COMBINATION OF NAMED ENTITY RECOGNITION TOOLS APPLIED TO BIOGRAPHIC TEXTS

ICSCS 2013 - Lille

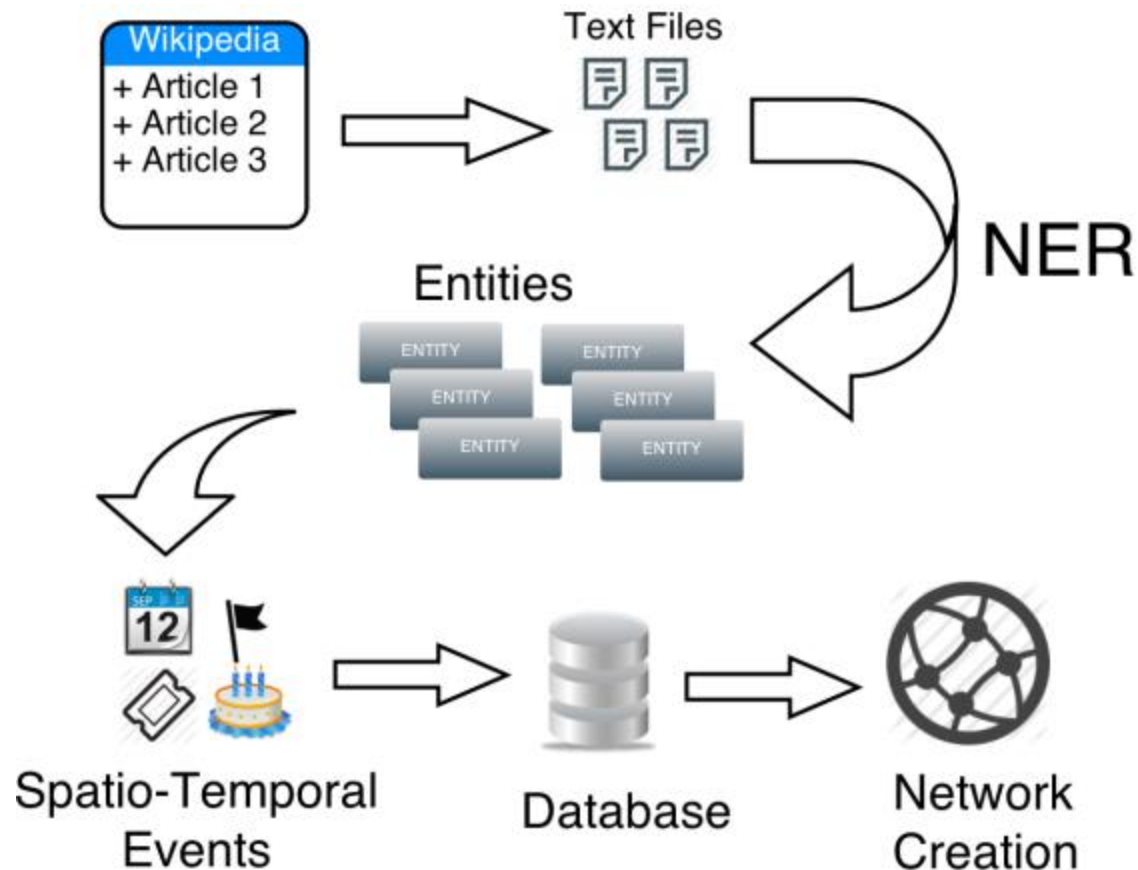
Samet Atdag
Vincent Labatut

Structure

- Problem
- Corpus
- NER Tools
- Comparison
- Results

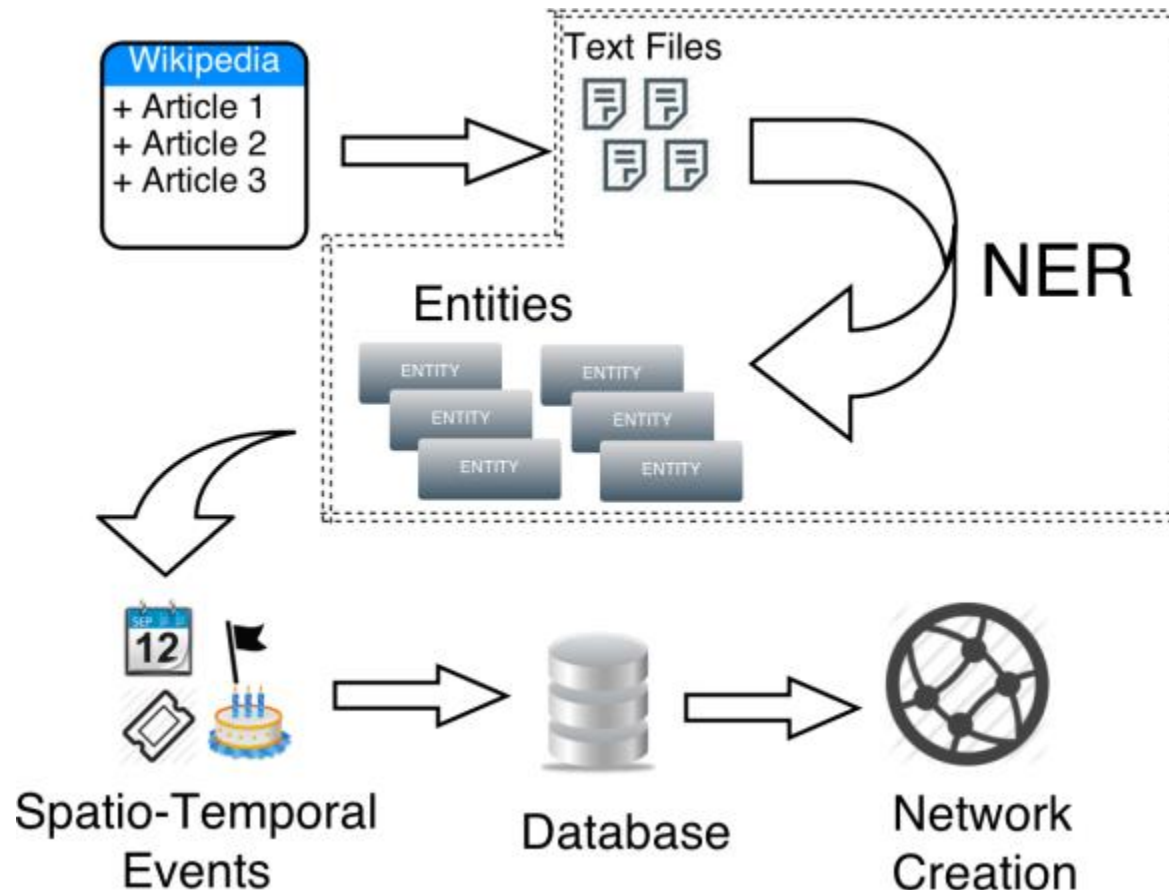
General Project

- A part of a long term project



General Project

- A part of a long term project



NER

- Definitions
 - Named Entity (NE)
 - Entity Type
 - Common entity types

Example:

John went to Lille.

Named entities: John, Lille

Entity types: person, location

Problem

- Long-term project constraints
 - Text: Biographic
 - Types: person, location, organization (date)
- Which NER tool?
 - Many NER tools.
 - Difficult to select

Comparison Method

1. Get appropriate data
2. Get relevant tools
3. Assess performance in details
4. Compare performance

Corpus – Review

- Existing corpora
- Many of them

Corpus	Domain
NYTAC	News articles
MUC	Military messages, terrorism reports, airplane crashes
NIST IE-ER 99	News
CoNLL03	News
Email Corpora	Emails
ACE2005	News

Corpus – Constitution

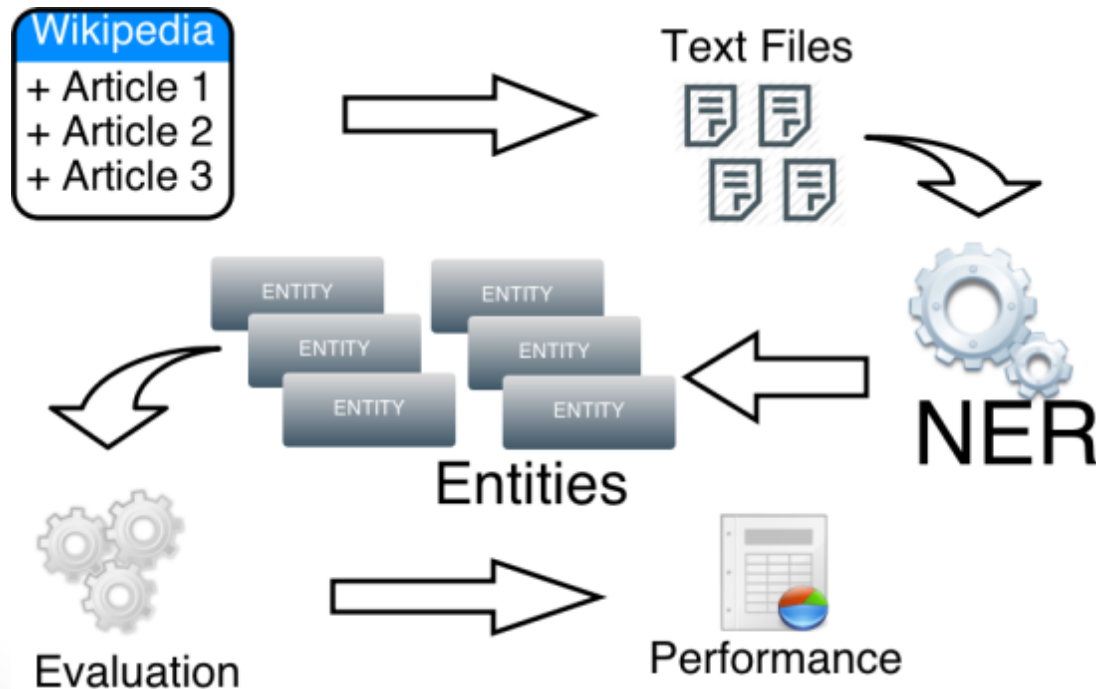
- None of them fit our purpose
 - Commercial, not available, domain specific
- Solution: Our corpus
 - Manually annotation before this project.
 - Size: 250 Wikipedia articles
 - Types: Person, Location, Organization(Date)
 - Critical problems:
 - Encoding
 - missing entities
 - not fully annotated
 - broken structure

NER Tools

- Selection criteria:
 - Freely available
 - Community supported
- 4 selected tools:
 - Stanford NER Tool
 - Illinois Named Entity Recognizer
 - OpenCalais Web Service
 - Alias-I Lingpipe
- External tools developed by different people/universities/initiatives

NERWIP

- Named Entity Recognition on Wikipedia
 - Initially developed as a license graduation project
 - Extensible; open for new NER tools
 - Extension for LingPipe
- Gets text, annotates, aggregates, evaluates.



Assessment – Classic Approach

- Classic approach: Using TruePositive, FalsePositive, FalseNegative
- Precision and Recall

$$Pre = \frac{TP}{TP + FP}$$

$$Rec = \frac{TP}{TP + FN}$$

- Simple, yet strong enough
- Gives normalized performance output.
- Important: combines types and spaces

Assessment – Example

Victor Charles Goldbloom was born in Montreal, the son of Alton Goldbloom and Annie Ballon. He studied at Selwyn House and Lower Canada College. He studied at McGill University receiving his BSc in 1944, his MD in 1945, his DipEd in 1950 and his DLitt in 1992. Dr. Goldbloom was assistant resident at the Columbia Presbyterian Medical Center, in New York.

- 10 actual entities (in boxes)
- 9 estimates (wavy underlines)
 - 5 TP (*Victor Charles Goldbloom, Montreal, Selwyn House, McGill University, New York*);
 - 4 FP (*Canada, MD, Dr. Goldbloom, Medical Center*)
 - 5 FN (*Alton Goldbloom, Annie Ballon, Lower Canada College, Goldbloom, Columbia Presbyterian Medical Center*).

Assessment – Our Approach

- For partial detections, modified Precision and Recall

Old Terms
-
False Negative
False Positive
True Positive

New Terms
Partial Match (PM)
Complete Miss (CM)
Wrong Hit (WH)
Full Match (FM)

- False Negatives = Partial Matches + Complete Misses
- False Positives = Partial Matches + Wrong Hits

Assessment – Our Approach (2)

Classic Approach:

$$Pre = \frac{TP}{TP + FP}$$

$$Rec = \frac{TP}{TP + FN}$$

Our approach:

- With using new terms,
4 new measures:
- Full Precision and
Partial Precision

$$Pre_F = \frac{FM}{FM + PM + WH}$$

$$Pre_P = \frac{PM}{FM + PM + WH}$$

- Full Recall and
Partial Recall

$$Rec_F = \frac{FM_i}{FM + PM + CM}$$

$$Rec_P = \frac{PM}{FM + PM + CM}$$

Assessment – Our Approach (2)

Classic Approach:

$$Pre = \frac{TP}{TP + FP}$$

$$Rec = \frac{TP}{TP + FN}$$

Our approach:

- With using new terms,
4 new measures:
- Full Precision and
Partial Precision

$$Pre_F = \frac{FM}{FM + PM + WH}$$
$$Pre_P = \frac{PM}{FM + PM + WH}$$

Replace FN with (PM+WH)

- Full Recall and
Partial Recall

$$Rec_F = \frac{FM_i}{FM + PM + CM}$$
$$Rec_P = \frac{PM}{FM + PM + CM}$$

Replace FP with (PM+CM)

Example – With new terms

Victor Charles Goldbloom was born in Montreal, the son of Alton Goldbloom and Annie Ballon. He studied at Selwyn House and Lower Canada College. He studied at McGill University receiving his BSc in 1944, his MD in 1945, his DipEd in 1950 and his DLitt in 1992. Dr. Goldbloom was assistant resident at the Columbia Presbyterian Medical Center, in New York.

- 5 FM (the entities previously considered as TP)
- 3 PM (*Lower Canada College, Dr. Goldbloom, Columbia Presbyterian Medical Center*)
- 1 WH (*MD*)
- 2 CM (*Alton Goldbloom and Annie Ballon*).

Results

Spatial and typical overall results for comparison

		Overall Spatial Performance
SNER	FPre	0.78
	PPre	0.10
	FRec	0.83
	PRec	0.10
INET	FPre	0.53
	PPre	0.26
	FRec	0.52
	PRec	0.26
OCWS	FPre	0.81
	PPre	0.10
	FRec	0.55
	PRec	0.06
LIPI	FPre	0.64
	PPre	0.17
	FRec	0.70
	PRec	0.19

		Overall Typical Performance
SNER	Pre.	0.88
	Rec.	0.93
INET	Pre.	0.80
	Rec.	0.78
OCWS	Pre.	0.91
	Rec.	0.61
LIPI	Pre.	0.82
	Rec.	0.88

Results

Spatial performance of OpenCalais

		Overall	Category					
			Art	Military	Politics	Science	Sports	Others
OCWS	FP	0.81	0.77	0.75	0.8	0.82	0.92	0.80
	PP	0.10	0.11	0.16	0.10	0.12	0.05	0.12
	FR	0.55	0.56	0.44	0.45	0.46	0.44	0.49
	PR	0.06	0.08	0.09	0.06	0.07	0.03	0.07

Typical performance of OpenCalais

		Overall	Type		
			Person	Location	Organization
OCWS	FP	0.81	0.87	0.8	0.74
	PP	0.1	0.07	0.08	0.14
	FR	0.55	0.56	0.52	0.54
	PR	0.06	0.04	0.05	0.1

Results

- Overall: clear hierarchy:
 - Stanford > LingPipe > OpenCalais > Illinois
- But in details: tool performance varies:
 - Entity type
 - Article category
- Interpretation:
 - A globally bad tool may be very good at a specific entity
 - A very good tool may be bad on specific article type
- Conclusion
 - Partial matches should be considered
 - Simple normalized performance assessment measure is necessary
 - Combining skills of tools may produce a better tool, we worked on combining tools and got better results.